Cell-States v1.0 (11.03.2019)

Stefanie Grosswendt[1,*], Helene Kretzmer[1,*], Zachary D. Smith[2,3,4,*], Abhishek Sampath Kumar[1], Sven Klages[1], Bernd Timmermann[1], Shankar Mukherji[5] and Alexander Meissner[1,2,3]

[1] Department of Genome Regulation, Max Planck Institute for Molecular Genetics, Berlin, Germany
[2] Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
[3] Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA
[4] Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts, USA
[5] Department of Physics, Washington University in St. Louis, St. Louis, Missouri

[*] These authors contributed equally to this work.
Correspondence: meissner@molgen.mpg.de

Cell-state nomenclature

To name cell-states, we searched localized gene expression data from whole mount in situ-hybridization (e.g. using the Gene eXpressionDatabase (GXD): http://www.informatics.jax.org/expression.shtml), transgenic reporters or single cell expression analysis. Our state nomenclature is intended to serve as an orientation, referring to cell types, anatomical structures or compartments within the embryo that are most compatible with the differential expression status of 712 informative marker genes (see below) and their emergence between Embryonic day (E)6.5 to E8.5. Greater heterogeneity in origin or cellular identity beyond these cell-states may be possible, but is beyond the resolution of our current clustering strategy.

0 - trophoblast stem cells	21 - primitive heart tube
1 - neural ectoderm anterior	22 - primitive blood late
2 - primitive streak late	23 - notochord
3 - anterior primitive streak	24 - fore/midbrain
4 - primitive/definitive endoderm	25 - distal extraembryonic ectoderm
5 - allantois	26 - neuromesodermal progenitor early
6 - secondary heart field/ splanchnic lateral plate	27 - primordial germ cells
	28 - differentiated trophoblasts
7 - gut endoderm	29 - visceral endoderm early
8 - ectoderm early 1	30 - presomitic mesoderm
9 - primitive blood early	31 - neuromesodermal progenitor late
10 - preplacodal ectoderm	32 - angioblasts
11 - neural ectoderm posterior	33 - neural crest
12 - posterior lateral plate mesoderm	34 - pharyngeal arch mesoderm
13 - hematopoietic/endothelial progenitors	35 - similar to neural crest
14 - parietal endoderm	36 - primitive blood progenitors
15 - amnion mesoderm early	37 - primitive streak early
16 - surface ectoderm	38 - node
17 - epiblast	39 - future spinal cord
18 - somites	40 - visceral endoderm late
19 - ectoderm early 2	41 - amnion mesoderm late
20 - splanchnic lateral plate/ anterior paraxial mesoderm	

Single cell to cell-state assignment

In the additional data file "CellStateKernels," we provide the expression profile of each of the 42 cell states identified in the mouse embryos deposited here. To assign single cells to the cell-states, the Euclidean distance of the normalized log-expression profile of a single cell (see Methods) to the kernels is calculated. The distance is based on the 712 marker genes, which can be found in the "CellStateKernels" file. The cell-state with the closest distance defines the cell-state of the respective single cell.

Cell-state identification

Epiblast

Cluster 17 represents the "epiblast," the cell state of the embryo proper mainly composed of pluripotent cells that are not part of the primitive streak. Its presence is restricted to our E6.5 timepoint and it displays the highest expression of Otx2 (PMID: 15201223), Zic2 (PMID: 15261827), Nodal (PMID: 9056778) and Fgf5 (PMID: 24131634). Expression of Fgf4 excludes the possibility of cluster 17 being of extraembryonic ecto- and endoderm nature (PMID: 1618140) and high expression of Utf1 argued against a mesodermal nature, as this gene was reported to be expressed in pluripotent cells and extraembryonic ectoderm but not expressed in primitive mesoderm (PMID: 9524124), which agrees with its comparatively reduced prevalence in states of the primitive streak (see below).

Embryonic mesoderm lineage

The primitive streak states (37, 3, 2) were characterized by the expression of T (PMID: 1821859), Evx1 (PMID: 1349539) and Lhx1 (PMID: 10328927). We assigned cluster 37 to "primitive streak early" as it is present at E7.0 and expresses Mesp1, which is detected in ingressing mesodermal cells as they pass through the primitive streak (PMID: 10393122) as well as Dll1, which was reported for the primitive streak and mesoderm at this stage (PMID: 7671806). Cluster 3 is maximal at E7.0 and annotated as "anterior primitive streak" based on the expression of Foxa2 (PMID: 8375339, PMID: 12351174) and Gsc (PMID: 9671576).

Cluster 2 emerges around E7.5 as cluster 37 diminishes and contains cells of the primitive streak and nascent mesoderm. We annotated this state "primitive streak late," in line with its high expression of Hoxb1, which extends along the primitive streak and is present in the mesoderm of the posterior embryo around E7.5 (PMID: 1983472). Moreover, the expression of Eomes is substantially reduced compared to the primitive streak states that are mainly present at E7.0 (state 37 and 3).

Cell states of the caudal epiblast (26, 30, 31) also expressed T, Evx1 and Lhx1. Their more posterior position within the embryo is supported by continued Hoxb1, reported to be expressed not only in the posterior mesoderm but also in the ectoderm anterior of and lateral to the primitive streak at E7.75 (PMID: 1983472). Cluster 26 (present around E7.5-8.0) and cluster 31 (present around E8.0-8.5) have a high fraction of cells that co-express T and Sox2, which is characteristic of neuromesodermal progenitors (NMPs). Thus, these states were annotated as "NMPs early" (cluster 26) and "NMPs late" (cluster 31) according to their temporal dynamics. In addition to their mesodermal signature, both states express the neuroectodermal genes Sox3, Sp8, Nkx1-2 and Olig3 (reported to be expressed in NMPs and derivatives, PMID: 28826820). State 26 (NMPs early) appears to give rise to state 31, at least partially, which might also contain more neuroectodermally-committed cells, as indicated by a slightly increased fraction of cells expressing of Pax3 (PMID: 28826820).

In contrast, cluster 30 emerges around E8.0-8.5, does not express Sox2, but instead expresses genes of the "presomitic mesoderm" such as Tbx6 (PMID: 8954725), Snai1 (PMID: 9671584), Dll1 (PMID: 9671584) and Nrarp (PMID: 19268448). Notably, the intermediate mesodermal marker Osr1 (PMID: 10473132) is also detected in a large fraction of state 30 cells, suggesting that the state may also

contain cells of this tissue. Transcriptional similarity (e.g. the high expression of Hoxb1, Dll1, T) suggests that state 30 is derived from state 2, though they may also proceed directly from NMPs. Interestingly, state 31 has a substantially larger fraction of cells expressing Cyp26a1 compared to state 30, which may indicate a more caudal position within the embryo (PMID: 11520679).
Cluster 12 is most prevalent around E7.5-8.5 and comprised of cells of the lateral plate mesoderm. It expresses Foxf1 (PMID: 11124112) and Isl1 (PMID: 14667410). Similar to state 30 (presomitic mesoderm), cluster 12 may also contain cells of the intermediate mesoderm as indicated by expression of Osr1 (PMID: 10473132). Expressed genes specific to the caudal part of the embryo, such as Cdx4 (PMID: 19906845) and Hoxd9 (PMID: 1676674), also supports a posterior position. From these data, we assigned this state "posterior lateral plate mesoderm", though it may contain mesoderm of other structures as well.
Cluster 20 is present around E7.5-8.0 and resembles "splanchnic lateral/anterior paraxial mesoderm" (PMID: 21498416). This state specifically expresses the anterior mesoderm specific Tbx1 (PMID: 8853987), which was described to be exclusively expressed in the in the anterior embryonic mesoderm, rostral to the node at E7.5. Furthermore, this state expresses: Tcf21, which is expressed within the first branchial arch at E8.0 (PMID: 9733105); Isl1, which is expressed at the earliest stages of cardiac development (PMID: 25174608); and Prdm1, which is expressed in splanchnic mesoderm (PMID: 12204275).
Cluster 34 also expresses several genes characteristic for the pharyngeal mesoderm (PMID: 21498416) and is present mainly around E8.5. We termed this state "pharyngeal arch mesoderm" by the expression of Tbx1 (PMID: 8853987, PMID: 21364285), Six1 (PMID: 21364285), Eya1 (PMID: 21364285) and Pax9 (PMID: 9732271), which are associated with this tissue as well as other cranial mesenchymal structures. Cluster 6, present at E7.5-8.5, was termed "secondary heart field/splanchnic lateral plate mesoderm" due to the expression of Isl1, a marker for the secondary heart field (SHF, PMID: 25174608; PMID: 14667410). It also displays a high expression Hand2, which is expressed in the lateral mesoderm and developing heart from E7.75 onwards (PMID: 8533092). The expression of Foxf1 and Irx3 support the notion that state 6 contains cells of the splanchnic (Foxf1) and also of the somatic (Irx3) lateral plate mesoderm (PMID: 11124112). Overall, this state might be composed of many cells that eventually contribute to the formation of the heart tube. Cluster 21, present around E8.0 and 8.5, is largely composed of cells of the "primitive heart tube" according to the expression of Myl7 (PMID: 11245568), Nkx2-5 (PMID: 12141429, PMID: 11336496), Hcn4 and Tbx5 (PMID: 8853987), Mef2c (PMID: 8026334) and Myocd. Hcn4 and Myocd have been reported to be rather specific for the FHF (PMID: 23974038, PMID: 25174608). The scarcity of known markers that are distinct for the First and Second Heart Fields makes it difficult to conclude with certainty whether 6 gives rise to state 21 or whether 21 develops from state 20 independently and in parallel. Nevertheless, higher Hoxb1 expression in state 6 compared to state 21 suggests that it is located posterior to the primary heart tube.
Cluster 18 begins to emerge at E7.5 and continues through E8.0-8.5. We termed this state "somites" as it expresses Ripply2, which is reported to be expressed at the site of somite formation (Somite 0-1, PMID: 18045842), as well as Tbx18 (PMID: 11118889) and Pax3 (PMID: 18644785), which are expressed throughout formed somites. Higher expression of Aldh1a2 (PMID: 17849458) and Foxc1 (PMID: 8375339) in this cell state compared to cluster 30 (presomitic mesoderm) indicates that this state preferentially contains cells of the somatic rather than the presomitic mesoderm.
The entire lineage of hematopoietic and endothelial cells was marked by the expression of Tal1. Cluster 13 is present at E7.0 and is thereby the first detected state expressing Tal1, Kdr1 and Etv2, characteristic for "hematopoietic and

endothelial progenitors" (PMID: 24052951). Cluster 36 likely arises from Cluster 13 and was termed "primitive blood progenitors." It is most prevalent around E7.5 and characterized by the expression of Hbb-bh1, Gata1, Nfe2 (PMID: 29311656), Klf1 (PMID: 16380451) and Runx1 (PMID: 10226014), genes known to demarcate the first wave of blood production in the embryo.
Cell states 9 and 22 appear to be "primitive blood early and late." They are prevalent at E8.0 and E8.5, respectively, and express marker genes mentioned for cluster 36. Within this lineage, we see a progressively decreasing fraction of Runx1+ cells from state 36 through 22, consistent with previous reports (PMID: 10226014). Cell state 32 also appears to arise from cells of state 13 at E7.5. We characterized these cells as "angioblasts" by the high expression of Pecam1 (PMID: 24052951, PMID: 24550118), Cdh5 (PMID: 24052951), Tek (PMID: 8187650) and Lyve1 (PMID: 27880904). This state also expresses Gja5 and Ephb4, an arterial and venous signature consistent with their endothelial identity (PMID: 21793101).
Cluster 38 and 23 represent cells of the axial mesoderm identified by the expression of Noto (PMID: 15533813), Shh (PMID: 8069909) and Foxa2 (PMID: 8375339). Cluster 38 emerges at E7.0 and was therefore annotated to be the "node," formed at the anterior most position of the posterior streak (state 3). Subsequently, cluster 23 emerges around E7.5 and persists into E8.5, suggesting that these are cells of the developing "notochord" that expands anteriorly from the node.

Extraembryonic mesoderm lineage
Cluster 15, present around E7.5-8.5, and cluster 41, present around E8.0-8.5, were annotated as the "amnion mesoderm early" and "amnion mesoderm late" respectively according to the expression of Foxf1 (PMID: 11124112) and Postn, which is highly specific (PMID: 22966238). Foxf1 is also specifically expressed in the "allantois" (cluster 5), another extraembryonic mesoderm cell state that emerges around E7.0 and then persists throughout the developmental window investigated here. The allantois expresses Tbx4 (PMID: 21932311, PMID: 8853987) and Tbx20 (PMID: 10940636), as well as posterior Hox genes Hoxa10, Hoxa11 and Hoxa3 that are very specific to this tissue (PMID: 22219351).

Germline
Cluster 27, of which a small number of cells are already present at E6.5, is unambiguously the primordial germ cell state ("PGC") according to the expression of Dppa3 and Prdm1 (PMID: 18583473), Prdm14 (PMID: 18622394) and Nanos3 (PMID: 20174582).

Extraembryonic and embryonic endoderm lineages
We identified endodermal cells according to their expression of Foxa2 (PMID: 22236333, PMID: 8306889) and Sox17 (PMID: 11973269). Cluster 4 is most prevalent at E6.5 but persists through E7.0, and was annotated as "primitive and definitive endoderm," with two distinct placements in our lineage tree (Fig. 2D). Cluster 4 represents a large fraction of E6.5 embryos (17%), suggesting an abundance of extraembryonic endoderm, though it likely also contains cells of the emerging embryonic endoderm. Separating these cells further is complicated by their highly similar gene expression signatures, which diverge later to become more obvious (see differential analysis of gene expression for E7.5 and E8.25 in PMID: 22236333). Cluster 29 and 40 represent "visceral endoderm early and late" as these two states emerge successively and both express Cubn and Amn (PMID: 20637190) as well as Amot and Slc39a8 (PMID: 17576135). Cluster 14 appears to be "parietal endoderm," which expresses Srgn (PMID: 11369593) and Thbd (PMID: 8681807) as well as Gkn2 and Pga5 (PMID: 28012457). This state was captured irregularly across embryonic replicates but detected consistently over development, likely due to technical variability that results from its location as the outermost extraembryonic

layer. Both the parietal and, to a lesser extent, visceral endoderm express Sox7, which distinguishes them from cluster 7 (PMID: 11973269). Cluster 7 is present around E7.5-8.5 and composed of embryonic endoderm cells. Because the main endodermal structure present at these developmental stages is the primitive gut, the state was termed “gut endoderm.” It is characterized by the absence or comparatively reduced expression of the above mentioned visceral and parietal endoderm markers, as well as by presence the definitive endoderm markers Pyy (PMID: 17683524), Foxa1, Cldn8 and Sorcs2 (PMID: 22236333). Whether this cell state directly emerges from state 3 or proceeds through some intermediate cells contained within state 4 is unclear.

Embryonic ectoderm lineage

Cluster 8 and cluster 19 are both already present at E6.5 and are generally similar in gene composition to epiblast. Cluster 19 becomes less prevalent towards E7.5, while cluster 8 peaks in abundance at E7.0 and comprises a substantial fraction of the embryo. Both clusters are enriched in Otx2, Zic2, Utf1 and Fgf5 albeit to a lesser extent than in the epiblast state 17. Additionally, expression of Sox2 and Pou3f1 indicate that cells of these states originate from the more anterior part of the epiblast (PMID: 24131634), which is in line with the substantially lower fraction of cells expressing markers of the primitive streak. Their ectodermal character is supported by the expression of Sox3 (PMID: 10446282) and a small number of cells (~10%) with detectable Six3 and Hesx1, which are both neural markers (PMID: 24131634). Thus, we termed both states “ectoderm early 1” (cluster 8) and “ectoderm early 2” (cluster 19). We placed these states jointly within the lineage tree to highlight the ambiguity of their direct developmental relationship: it is unclear whether one differentiates into another or if they represent distinct regions of the differentiating pluripotent field. However, state 8 displays a higher fraction of cells expressing the epiblast markers, which indicates a more direct relationship to state 17. Furthermore, state 3 cells (anterior primitive streak) are transcriptionally most similar to state 19, suggesting that state 19 may be comprised of cells that are localized in a similar region within the epiblast or are fated to transverse through the anterior primitive streak.

Cluster 1, 11, 24, and 39 represent neuroectodermal clusters characterized by the expression of Sox1, Sox2 (which we consider to be a neuroectodermal marker from E7.5 onwards) and Sox3 (PMID: 10446282). Cluster 24 emerges at E8.0 and expresses fore- and midbrain-associated Otx2, the midbrain marker Wnt1, as well as En1 and Pax2, which mark the mid-hindbrain border (PMID: 11253000). This cluster also contains cells that express Foxg1, which emerges around E8.5 within the anterior neural folds (PMID: 16530751) and Six3, reported for the most rostral neuroectoderm (PMID: 11532921). Taken together, we conclude that cluster 24 represents the developing “fore/midbrain” and likely emerges from cluster 1, which is present at E7.5-8.0 and is Six3, Otx2, Pax2 and En1 positive. Consequently, cluster 1 was termed “neural ectoderm anterior.” Cluster 11 is most prevalent at E7.5 and annotated as “neural ectoderm posterior” due to the expression of the hindbrain marker Gbx2 (PMID: 11253000, PMID: 11532921) and Hoxa1 (PMID: 9053316). Cluster 39 is present at E8.0-8.5 and appears to be “future spinal cord” by its expression of Hoxb8 (PMID: 8096483), Nkx6-1 and Olig2 (PMID: 15652703). We placed this cell state downstream of state 11 because they both represent posterior neuroectodermal cells and emerge successively. However, we include a dashed line in our lineage tree to account for a possible dual origin from neuromesodermal progenitor cell states (26 and 31).

Cluster 33 is most prevalent at E8.5 and appears to be “neural crest” by the expression of Sox10, Sox9, Twist, Foxd3, and Tfap2a (PMID: 24780627, PMID: 22889333, PMID: 22889333). It was placed in the tree downstream of the developing brain (state 24). However, more posterior structures like the spinal cord also produce

neural crest cells and are expected to contribute to this cell state. Cluster 35 is present around the same time and did not display any enrichment of specific marker genes that would have allowed an anatomical or cell type specific annotation. Its transcriptional profile appears most similar to state 33, which is why it was termed "similar to neural crest" and place adjacent to it in the lineage tree.
Cluster 10 and 16 both emerge around E7.5 and progressively increase in relative abundance. The presence of Dlx5 suggests that both are non-neural ectoderm (PMID: 9763476). Cluster 10 expresses Six1 and Eya1, consistent with a "preplacodal ectoderm" identity (PMID: 19027001), whereas cluster 16 was annotated as "surface ectoderm" because it expresses Trp63 (PMID: 14757276), Tfap2c (PMID: 1989904) and Grhl3 (PMID: 16831572).

Extraembryonic ectoderm lineage
Cluster 0, 25 and 28 are Elf5 positive and belong within the extraembryonic ectoderm (ExE) lineage (PMID: 25446535). Cells of cluster 0 display the highest expression of Eomes, Cdx2, Sox2 and Esrrb and are annotated as "trophoblast stem cells" (PMID: 25446535, PMID: 11433360). Along with Spry4 (PMID: 25446535), the high expression of these genes also suggests that cluster 0 is located proximal to the epiblast. In contrast, canonical trophoblast stem cell markers are detected in fewer cluster 25 cells, which we term "distal ExE." This decrease suggests that cells are undergoing differentiation or that they are located more distal to the epiblast; both observations are in line with an enrichment for Ets2 and Ascl2 expression in this state as previously reported for the distal ExE/ectoplacental cone (PMID: 25446535). Cluster 28 shows substantially lower frequencies of trophoblast stem-cell markers and expresses differentiation markers like Plac1 and Prl3d1, indicating that it is comprised of "differentiated trophoblasts," including cells of the ectoplacental cone and trophoblast giant cells (PMID: 25446535, PMID: 18662396).

Computational Methods

Preprocessing
The Cell Ranger pipeline (10X Genomics Inc.) was used for each scRNA-seq data set to de-multiplex the raw base call files, generate the fastq files, perform the alignment against the mouse reference genome mm10, filter the alignment and count barcodes and UMIs. Outputs from multiple sequencing runs were also combined using Cell Ranger functions. Furthermore, single cells were assigned to embryos according to the autosomal fraction of CAST SNPs.

Cluster determination
The cluster determination was split into three main parts and was largely done using the R package Seurat with default settings. In brief, (1) A preliminary set of clusters were generated by agnostically clustering WT embryos of the same stage as a pool without taking replicate identity into account, followed by generating per replicate clusters according to this assignment. Then, (2) replicate embryo clusters from step 1 were clustered across time points to obtain preliminary cell states. Finally, (3) all cells were assigned to their most similar cluster by Euclidean distance according to a reduced set of 712 marker genes to determine their specific cell state identity.

(1) Embryo specific centers (WT): All wild type single cells of the same developmental stage were processed together after discarding cells that were not confidently assigned to a genotype/embryo. Parameters were adopted from the Seurat manual. The expression data were log-normalized, scaled to 10,000 and UMI biases were removed (vars.to.regress = "nUMI"), followed by calling of variable genes (parameters: mean.function = ExpMean, dispersion.function = LogVMR,

x.low.cutoff = 0.0125, x.high.cutoff = 3, y.cutoff = 0.5). Next, the variable genes were used to run the PCA and the first 20 PC's were used for cluster detection. The average expression for each embryo and cluster was calculated, which we refer to as "embryo-specific centers." This allowed us to detect even rare cell states while preserving embryo-specific variability.

(2) Cell cluster (WT): The individual embryo specific centers were combined from all stages into one analysis to determine variable genes. A PCA was run based on the variable genes and the first 20 PC's were used to cluster the embryo specific centers (parameters adjusted for low 'cell' number: k.param = 8, k.scale = 50, prune.SNN = 1/10). This resulted in 42 clusters defined by the median expression profile from each contributing embryo-specific center, which we used as a preliminary cell state description. Then, as a temporary step, all WT cells from all stages were simultaneously assigned to their closest preliminary cell states based on expression similarity (Euclidean distance of log-expression values of variable genes calculated above) to calculate a gene expression average. At this stage, we observed that the number of variable genes was unevenly distributed across preliminary cell states, which created biases when comparing single cells across them (clusters defined by a greater number of variable genes have more opportunities to match sparse single cell measurements, while those defined by fewer variable genes accumulate more noise by including them). We therefore sought to normalize the number of state-specific genes that contribute to each cluster by using the top 30 marker genes (highest difference in fraction of positive cells within the cluster versus other clusters) from each of the 42 cell states. We found that this reduced gene set provides a more stable, lower-noise assignment without biasing the information to describe each cell state (n = 712 unique genes).

(3) Cell states of single cells: The WT cells were assigned to the cell states based on their Euclidean distance log-expression values for the 712 marker genes. Single cell distances were found to be significantly smaller to their matched cell states than to next-best matches.